ORIGINAL REPORT

TEST-RETEST AND ALTERNATE FORMS RELIABILITY OF THE ASSISTING HAND ASSESSMENT

Marie Holmefur, OT, PhD[1,2], Pauline Aarts, OT, MSc[3], Brian Hoare, OT[4] and Lena Krumlinde-Sundholm, OT, PhD[1]

From the [1]Neuropediatric Research Unit, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, [2]Centre for Rehabilitation Research, Örebro County Council, Örebro, Sweden, [3]Department of Child Rehabilitation, Sint Maartenskliniek Nijmegen, The Netherlands and [4]Victorian Paediatric Rehabilitation Service, Monash Medical Centre, Melbourne, Australia

Objective: **The Assisting Hand Assessment (AHA) has earlier demonstrated excellent validity and rater reliability. This study aimed to evaluate test-retest reliability of the AHA and alternate forms reliability between Small kids vs School kids AHA and the 2 board games in School kids AHA.**

Design: **Test-retest and alternate forms reliability was evaluated by repeated testing with 2 weeks interval.**

Subjects: **Fifty-five children with unilateral cerebral palsy, age range 2 years and 3 months to 11 years and 2 months.**

Methods: **Intraclass correlation coefficients and smallest detectable difference were calculated. Common item and common person linking plots using Rasch analysis and Bland-Altman plots were created.**

Results: **Intraclass correlation coefficients for test-retest was 0.99. Alternate forms intraclass correlation coefficients were 0.99 between Small kids and School kids AHA and 0.98 between board games. Smallest detectable difference was 3.89 points (sum scores). Items in common item linking plots and persons in common person linking plots were within 95% confidence intervals, indicating equivalence across test forms.**

Conclusion: **The AHA has excellent test-retest and alternate forms reliability. A change of 4 points or more between test occasions represents a significant change. Different forms of the AHA give equivalent results.**

Key words: reliability and validity, outcome assessment, cerebral palsy, hand function, Assisting Hand Assessment.

INTRODUCTION

The number of outcome measures or tests available in children's rehabilitation has grown rapidly (1). Measures are fundamental tools both within research and clinical practice, and form the basis for evidence-based practice. Some requirements have to be met for a test to be useful. Most importantly, a test needs to be valid, which refers to whether the test is measuring what it is intended to measure. Measures are not valid unless they have high precision and demonstrate good reliability (2). This study is one step in a process of evaluating the psychometric properties of the Assisting Hand Assessment[1] (AHA). The purpose of the AHA is to measure and describe how effectively children with unilateral disabilities use their affected hand when performing bimanual activities. The AHA was developed for children with unilateral cerebral palsy and obstetric brachial plexus palsy (OBPP) aged 18 months to 12 years of age (3, 4). Administration of the AHA involves a play session using bimanual activities, which serve as a means for observing a child's typical way of using his or her affected hand. For children aged 18 months to 5 years the play session involves spontaneous play with toys requiring 2 hands (Small kids AHA). For children aged 6–12 years a recently developed board game (School kids AHA) provides an age-appropriate context for handling the same toys as in the Small kids AHA. In order to enhance motivation and flexibility, the board game has 2 different contexts; firstly, the Alien game, set in space, and secondly, a fantasy theme, called the Fortress game. It is recommended to use the board games for children from the age of 6 years.

The psychometric properties of the AHA have been described in earlier studies (3–6). The validity and aspects of reliability were evaluated using Rasch measurement analysis with excellent results. It was shown that AHA items measure a unidimensional construct, has very good targeting of item difficulties to person abilities and the scale has a high person separation measure (3–5). Inter-rater and intra-rater reliability of the Small kids AHA was investigated and found to be excellent. Intraclass correlation coefficient (ICC) for inter-rater was 0.97 in a 20-rater design and 0.98 in a 2-rater design and 0.99 for intra-rater reliability (6).

There are a few aspects of reliability remaining to be evaluated for the AHA, one of which is test-retest reliability. In all measures a certain degree of variation is present and a test-retest evaluation reveals the magnitude of measurement error caused

---

[1]Courses and test material are provided by Handfast AB, Stockholm, Sweden.

by chance, rater variation or variation of the performance of the assessed person on different occasions. With established test-retest reliability it is possible to estimate whether and to what extent the possible differences found on the measure are due to a real change in the person's ability or are within the measurement error of the test. Test-retest reliability is evaluated by testing subjects on repeated occasions; the measure is repeated in a time interval short enough for the ability to be stable and long enough to avoid learning or memory effects (7). The term intra-rater is sometimes used interchangeably with test-retest. Here a difference is made between the 2: intra-rater reliability is the agreement between repeated observations of the same test session (i.e. using a videotape). This isolates intra-rater error, since the performance of the individual on the videotape does not change. When evaluating test-retest reliability the individual is tested twice, in the case of AHA this involves repeated video-taping. Test-retest reliability inevitably includes intra-rater error. The other aspect of reliability remaining to be evaluated for the AHA is whether the different forms of the test (Small kids vs School kids and the Alien game vs the Fortress game) give equal results. One purpose of developing the AHA for older ages was to be able to measure development over time, and the 2 forms of the board game were developed to be used interchangeably. In both cases we need to know whether the test scores obtained from different forms of the test are comparable. Therefore the alternate forms reliability (2, 8) needs to be evaluated for the AHA. Thus, the aim of this study was to evaluate test-retest reliability of the AHA and to evaluate alternate forms reliability between the Small kids vs School kids AHA and between the Alien vs Fortress forms of School kids AHA.

METHODS

Participants

A convenience sample of 55 children, age range 2 years and 3 months to 11 years and 2 months (mean 5 years and 8 months, standard deviation (SD) 26 months), divided into 3 groups were included in the study (Table I). The aim was to recruit 18 children in each group, with sample size calculated using the method by Walter et al. (9). The calculation was based on the assumptions that the lowest acceptable ICC was 0.7 and the target ICC was 0.9. The children were recruited via outpatient rehabilitation clinics and a special school setting in Australia, the Netherlands and Sweden. All children were diagnosed with unilateral cerebral palsy. No children received any form of intensive therapy to improve hand function between test sessions.

Informed consent was obtained from both children and parents. The study was approved by the Human Research Ethics Committee of Southern Health in Melbourne, Australia, the regional Medical Ethical Committee on Research Involving Human Subjects, Arnhem-Nijmegen, The Netherlands and the Regional Ethics Board at Karolinska Institutet, Stockholm, Sweden.

Measure

In an AHA assessment a video-recorded play session is conducted in a standardized manner (10). Using the video recording the child's use of the affected hand is assessed on 22 items using a 4-point rating scale. The same items and scoring criteria are used for all forms of the AHA. A raw sum score is produced ranging from 22 to 88 points, with higher score indicating better ability. Using Rasch analysis, the sum scores are converted to a logit measure, with a range from –10.18 to 8.70 logits. For this study, logit measures were obtained in a Rasch analysis using anchored values for both item measures and item structure from the data set used in Krumlinde-Sundholm et al. (4). The AHA was administered and scored by certified raters.

Design and procedure

This study had 3 groups: group 1 ($n = 18$) evaluated test-retest reliability of Small kids AHA, aged 18 months to 5 years; group 2 ($n = 18$) evaluated alternate forms reliability of Small kids vs School kids AHA, aged 5–6 years. Half of the group was allocated to the Small kids AHA for the first session and the other half to the School kids AHA. Both board games were used; and group 3 ($n = 19$) evaluated alternate forms reliability of the Alien vs Fortress board games in the School kids AHA, aged 6–12 years. Ten children were allocated to the Fortress game and 9 to the Alien game for the first session and groups were then swapped.

All children were filmed twice, with approximately 2 week's interval (mean 13 days, SD 5 days). All play sessions were assessed by the same rater to avoid involvement of inter-rater bias. To reduce potential rater bias due to memory of a child's first assessment whilst rating their second assessment, both videos of one child were not assessed during the same day and there was always assessment of other children undertaken in between ratings of the same child.

Statistical analysis

Reliability is usually expressed as a reliability coefficient, which estimates to what extent test scores are free from measurement error (7, 8). In this study the ICC was calculated using a one-way analysis of variance (ANOVA) (model 1,1 according to Shrout & Fleiss (11)) as the reliability coefficient for both sum scores and logit measures. To analyse whether there were systematic differences between test occasions and test forms Bland-Altman plots were created. In the Bland-Altman plot differences between test sessions were plotted against their mean and the limits of agreement were calculated as the mean difference ± 2 SD of the difference (12). To evaluate reliability on an item level both ICC (1,1) and percentage agreement were calculated. For items with ICC below 0.70 Wilcoxon signed-ranks test was conducted to analyse whether the low ICC was due to systematic differences between sessions or forms.

The standard error of measurement (SEM) gives clinically useful information as it expresses measurement error in the same unit as the test measure. The SEM was calculated from an ANOVA analysis as the square root of the pooled MeanSquare-time and MeanSquare-person × time. The SEM can be used to calculate a 95% confidence interval of ± 1.96 SEM around the sum score (13). The smallest detectable difference (SDD) expresses the smallest change that must take place between 2 measurements for the test to detect a real change with 95% certainty (14, 15). The SDD was calculated from the SEM; SDD = SEM × 1.96 × √2 (14). The SEM and subsequently SDD were calculated for AHA sum score, both in the unit sum scores and logits. The level of significance for the resulting SDD was calculated from the formula $Z = (\text{measure 1} - \text{measure 2})/\sqrt{2}$ SEM, as outlined by Eliasziw et al. (16).

Alternate forms of the AHA were further compared in the context of Rasch analysis. The item difficulties were compared in a common item

Table I. *Participants; age at first filming, sex and affected side in each study group (n = 55)*

Group	*n*	Age range	Age, mean (SD)	Sex, girls/ boys, *n*	Affected side of body, right/left, *n*
1	18	2 y 3 m–4 y 11 m	3 y 3 m (8 m)	6/12	10/8
2	18	5 y 1 m–6 y 11 m	5 y 8 m (6 m)	9/9	12/6
3	19	6 y 1 m–11 y 2 m	7 y 10 m (20 m)	6/13	11/8

m: months; SD: standard deviation; y: years.

Table II. *Intraclass correlation coefficients (ICC), standard error of measurement (SEM) and smallest detectable difference (SDD) for test-retest and alternate forms reliability of the Assisting Hand Assessment (AHA)*

| | | SEM | | SDD | |
	ICC (95% CI)[a]	Sum scores	Logits	Sum scores	Logits
Test-retest reliability					
Small kids (*n*=18)	0.99 (0.97–0.99)	1.40	0.35	3.89	0.97
Alternate forms reliability					
Small kids vs School kids (*n*=18)	0.99 (0.98–0.99)	1.15	0.25	3.19	0.68
Alien vs Fortress (*n*=19)	0.98 (0.96–0.99)	1.32	0.28	3.65	0.76

[a]ICC was identical for sum scores and for logit scores of the AHA.
CI: confidence interval.

linking plot. This was done by plotting item difficulties for 2 forms against each other with a 95% confidence band drawn in the plot. If 95% of the items fell within the confidence bands the test item estimates were regarded as invariant (17). Person ability estimates from different test forms were plotted in a common person linking plot. The 2 forms were regarded as measuring the same construct if 95% of the person measures fell within the confidence bands. The relative difficulty of the tests is shown by the position of the empirical line crossing the y-axis. If the empirical line falls through the origin the forms can be said to have equal difficulty (17).

RESULTS

Test-retest ICC was 0.99 for the Small kids AHA. Alternate forms ICCs were 0.99 between the Small kids and the School kids AHA and 0.98 between the Alien and the Fortress board games in the School kids AHA (Tables II and III). Reported ICCs are valid both for sum scores and logit measures. Bland-Altman plots show that all differences between sessions are within the limits of agreement (Fig. 1).

The SEM from test-retest data on AHA was 1.40 for the Small kids AHA, giving a SDD of 3.89 (Table II). A change of 3.89 sum scores has a significance of $p=0.049$.

Item ICCs and percentage of total agreement for both test-retest and alternate forms are shown in Table IV. In test-retest for Small kids 18 of 22 items have ICCs over 0.70. Of the

Table III. *Mean and standard deviation (SD) for each test session or test form in the 3 parts of the reliability study*

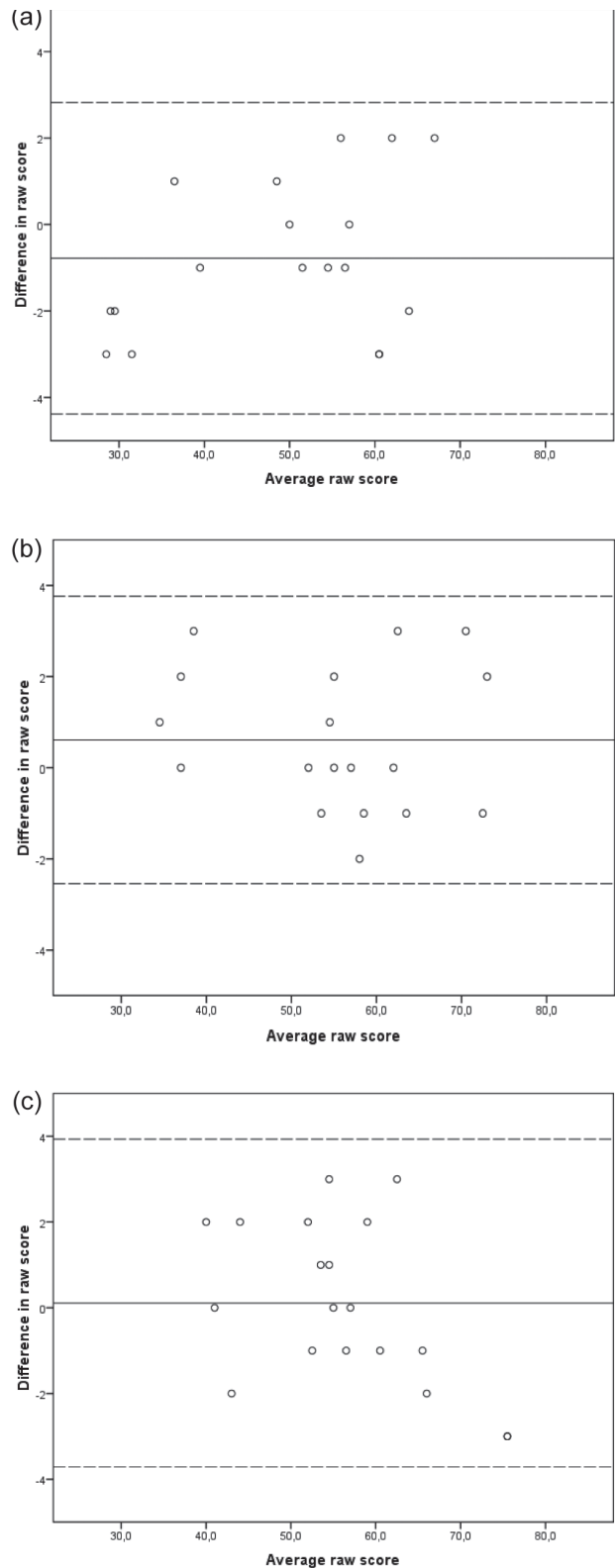|	AHA sum scores mean (SD)	AHA logits mean (SD)
Test-retest reliability		
Test	48.6 (13.6)	−1.23 (3.13)
Retest	49.4 (12.8)	−1.01 (2.91)
Alternate forms reliability		
Small kids vs School kids		
Small kids	55.6 (11.9)	0.34 (2.59)
School kids	54.9 (12.1)	0.20 (2.67)
Alien vs Fortress		
Alien	56.3 (9.7)	0.50 (2.11)
Fortress	56.2 (10.7)	0.44 (2.26)

AHA: Assisting Hand Assessment.

Fig. 1. Bland-Altman plots: difference against mean for Assisting Hand Assessment (AHA): (a) test-retest for Small kids AHA; (b) alternate forms between Small kids vs School kids; and (c) alternate forms between Alien and Fortress forms of the School kids AHA. Solid line: group mean difference. Dotted line: limits of agreement according to Bland & Altman.

Table IV. *Test-retest and alternate forms reliability of individual items*

	Test-retest reliability Small kids		Alternate forms reliability			
			Small kids vs School kids		Alien vs Fortress	
	ICC	Total agreement, %	ICC	Total agreement, %	ICC	Total agreement, %
General use items						
Approaches objects	0.80	89	1.00	100	1.00	100
Initiates use	0.75	72	0.83	89	0.59	74
Chooses assisting hand when closer to objects	na[1]	100	1.00	100	−0.03	89
Arm use items						
Stabilizes by weight or support	0.98	94	1.00	100	1.00	100
Reaches	0.75	78	0.95	94	0.82	68
Moves upper arm	0.12	61	0.88	89	0.92	95
Moves forearm	0.91	94	1.00	100	0.92	95
Grasp – release items						
Grasps	1.00	100	1.00	100	1.00	100
Holds	0.94	83	0.83	94	1.00	100
Stabilizes by grip	0.93	89	0.94	89	0.96	95
Readjusts grip	0.94	94	0.93	89	0.80	68
Varies type of grasp	0.86	83	0.97	94	0.75	68
Releases	0.94	89	0.94	83	0.89	89
Puts down	0.00	94	1.00	100	0.65	95
Fine motor adjustment items						
Moves fingers	0.82	83	0.91	89	0.90	95
Calibrates	0.91	89	0.96	94	0.90	84
Manipulates	0.79	89	0.76	89	0.59	79
Coordination items						
Coordinates	0.92	83	0.93	89	0.83	84
Orients objects	0.94	89	0.94	89	0.87	79
Pace items						
Proceeds	0.80	78	0.64	67	0.80	74
Changes strategy	0.55	61	0.57	72	0.85	79
Flow in bimanual task performance	0.93	94	1.00	100	0.92	95

[1]The ICC could not be calculated since all persons were given the score 2 and thus no variance was present for the item.

ICC: intraclass correlation coefficients; na: not applicable.

4 items with lower ICCs, 2 can be explained by low variance, as they have 100% and 94% total agreement. The remaining items "changes strategies" and "moves upper arm" have both low ICC and total agreement. There was however no systematic difference between test and test-retest for "changes strategies", $p = 0.257$, and "moves upper arm", $p = 0.206$. In the alternate forms comparison of Small kids and School kids, the item "proceeds" had low ICC and total agreement, but there was no significant difference between the alternate forms ($p = 0.102$).

In alternate forms for School kids 4 items had ICCs lower than 0.70; however, they all had total agreement over 70%.

Common item linking plots for both alternate forms trials (Fig. 2) show that all items in each plot are within the confidence bands. Thus, as a whole test items are equally difficult between forms of the AHA. In the Small kids vs School kids plot, the "changes strategies" item falls very close to the confidence band; indicating that the School kid's activity may place slightly higher demands on children than the Small kid's activity for this
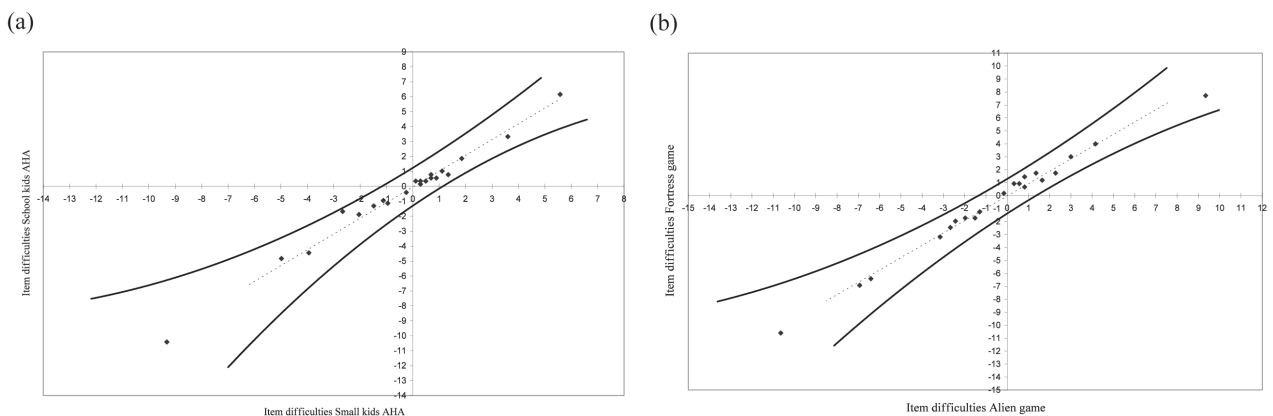
(a)

(b)

Fig. 2. Common item linking plots: (a) Small kids vs School kids Assisting Hand Assessment (AHA); (b) Alien vs Fortress games.
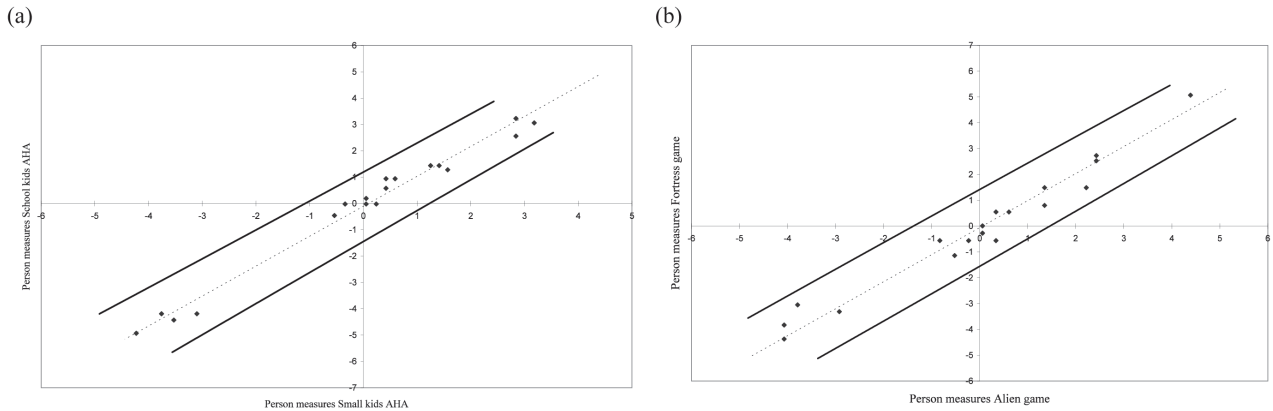
(a)

(b)

Fig. 3. Common person linking plots: (a) Small kids vs School kids Assisting Hand Assessment (AHA); (b) Alien vs Fortress games.

item. In the common person linking plots (Fig. 3), all person measures fall within the confidence bands, indicating that the compared forms measure the same construct. Also in both plots the empirical line falls through the origin, meaning that the relative difficulties of the compared forms are equal.

DISCUSSION

This study reports evidence of excellent test-retest reliability for the Small kids AHA. The high ICC indicates that the AHA is a stable test, with children's behaviour being very similar in repeated test sessions. The test-retest-SEM of 1.4 points (sum score) can be compared to the intra-rater SEM published earlier, which was 1.2 points (6). Considering test-retest error of the AHA consists of both intra-rater error, variation by chance and between-session variability in the child, the difference between the 2 SEMs is remarkably low. This indicates test-retest variability is, to a large extent, due to intra-rater error, and children's performance is very stable across test sessions.

The SDD for test-retest of 3.89 sum scores indicates that a change in AHA scores from one test session to the next must be 4 sum scores or more to be considered a true change with 95% certainty. Eliasziw et al. (16) suggested calculating the significance level for a given score change. When using their formula (see Methods section) a score change of 4 sum scores is a significant change with $p = 0.046$.

The test-retest reliability for the School kids AHA was not explicitly evaluated in this study because a "pure" test-retest trial would have involved repeated testing with the same board game. Such a trial would logically involve less error than our alternate forms trial. Some conclusions about the test-retest reliability for this age group can, however, be drawn from the alternate forms trial, where the children were tested with different board games with 2 weeks interval. From this follows that the alternate forms ICC of 0.98 that was found between the board games is the minimum test-retest ICC, thus, test-retest reliability can be assumed to be excellent also for School kids AHA. Likewise, the SDD of 3.65 is the maximum SDD for the School kids AHA, and thus a change of 4 sum scores or more can be considered a true change.

This study demonstrates excellent alternate forms reliability of the Small kids vs School kids AHA, and the Alien vs Fortress board games. From the results in the alternate forms trial we conclude that testing with both the Small kids and School kids AHA and the 2 different board games give directly comparable results. The fact that the SDDs for alternate forms were not higher than for the test-retest design indicates that changing between test forms as children grow older does not increase the variability in test results. Thus, when using different forms of the AHA the SDD of 4 sum scores is still valid.

Test-retest reliability of the AHA has earlier been investigated in 2 studies by Buffart and colleagues (18, 19) involving children with both congenital transverse reduction deficiency (with and without prosthesis) and radius deficiencies. They reported lower ICC values, varying between 0.70 and 0.94. These results are not easily compared with ours, particularly due to the lack of validation for use of the AHA in children with unilateral upper limb reduction deficiencies. Using the current items may lead to difficulty in interpreting the manual resulting in unreliable scores. Adjustment of items and validation of the AHA for children with upper limb reduction deficiency is currently being undertaken.

The AHA was earlier validated for both children with unilateral cerebral palsy and OBPP, but in this trial no children with OBPP were recruited, which is a limitation of the study. This was mainly due to practical reasons. At recruitment sites there were very few children with OBPP. Therefore, a decision to conduct the study without this group represented was made. In children with cerebral palsy there is evidence of more variability in testing than other populations, e.g. measures of range of movement may vary between occasions due to fluctuating muscle tone (20). This is also commonly known and taken into account in experimental studies (21). Because children with OBPP do not have a central nervous system damage it is likely that they will be more consistent in their performance. Despite these limitations, it is most likely that the SDD found is also valid for children with OBPP.

Another limitation in this study was that there were no children under the age of 2 years in the test-retest trial. It is quite possible that children as young as 18 months display

more variability across test occasions; therefore it would be interesting to conduct a trial including these very young children to compare results.

In this study a range of statistical methods were used to evaluate data. This is in many ways consistent with Lexell & Downham's (22) proposal for reliability studies. We found that these different statistics complemented each other. For example, the use of percentage agreement is not usually recommended (8); however, we found it useful for providing additional information on the item ICCs, an analysis that is very dependent on variance. The ICC is commonly reported, but merely informs about the magnitude of measurement error relative to between-subject variability for the group tested. Therefore we find the SEM and SDD to be more useful, both in clinical practice and research, to inform about the measurement error for the scores of an individual and the amount of change that constitutes a real change.

In conclusion, this study has shown that test-retest and alternate forms reliability of the AHA is excellent. Test scores from the Small kids AHA and School kids AHA, as well as the 2 board games in the School kids AHA are directly comparable. Thus, reliable AHA-measures can be produced for children over the age span 18 months to 12 years by use of the different test versions. A score change on the AHA of 4 sum scores or more between 2 test sessions is, with 95% certainty, a real change.

REFERENCES

1. Law M, King G, Russell D, MacKinnon E, Hurley P, Murphy C. Measuring outcomes in children's rehabilitation: a decision protocol. Arch Phys Med Rehabil 1999; 80: 629–636.
2. American Educational Research Association APANCoMiE. Standards for educational and psychological testing. Washington: American Educational Research Association; 1999.
3. Krumlinde-Sundholm L, Eliasson A-C. Development of the Assisting Hand Assessment: a Rasch-built measure intended for children with unilateral upper limb impairments. Scand J Occup Ther 2003; 10: 16–26.
4. Krumlinde-Sundholm L, Holmefur M, Kottorp A, Eliasson A-C. The Assisting Hand Assessment: current evidence of validity, reliability and responsiveness to change. Dev Med Child Neurol 2007; 49: 259–264.
5. Holmefur M. The Assisting Hand Assessment: continued development, psychometrics and longitudinal use. PhD thesis. Stockholm: Karolinska Institutet; 2009.
6. Holmefur M, Krumlinde-Sundholm L, Eliasson A-C. Interrater and intrarater reliability of the Assisting Hand Assessment. Am J Occup Ther 2007; 61: 79–84.
7. Streiner D, Norman G. Health measurement scales. 3rd edn. Oxford: Oxford University Press, 2003.
8. Portney L, Watkins M. Foundations of clinical research; applications for practice. 2nd ed. Upper Saddle River, NJ: Prentice-Hall, Inc.; 2000.
9. Walter S, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998; 17: 101–110.
10. Krumlinde-Sundholm L, Holmefur M, Eliasson A-C. Manual: the Assisting Hand Assessment, version 4.4. Stockholm: 2007.
11. Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420–428.
12. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 327: 307–310.
13. Thorndike R. Reliability. In: Bolton B, editor. Handbook of measurement and evaluation in rehabilitation. Baltimore: Paul H. Brookes Publishing Co.; 1987, p. 21–36.
14. Schreuders T, Roebroeck M, Goumans J, van Nieuwenhuijzen J, Stijnen T, Stam H. Measurement error in grip and pinch force measurements in patients with hand injuries. Phys Ther 2003; 83: 806–814.
15. Beckerman H, Roebroeck M, Lankhorst G, Becher J, Bezemer P, Verbeek A. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001; 10: 571–578.
16. Eliasziw M, Young L, Woodbury G, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther 1994; 74: 777–788.
17. Bond TG, Fox CM. Applying the Rasch Model: fundamental measurement in the human sciences. 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2007.
18. Buffart L, Roebroeck M, Janssen W, Hoekstra A, Hovius S, Stam H. Comparison of instruments to assess hand function in children with radius deficiencies. J Hand Surg Am 2007; 32A: 531–540.
19. Buffart L, Roebroeck M, Heijningen V, Pesch-Batenburg J, Stam H. Evaluation of arm and prosthetic functioning in children with a congenital transverse reduction deficiency of the upper limb. J Rehabil Med 2007; 39: 379–386.
20. McDowell B, Hewitt V, Nurse A, Weston T, Baker R. The variability of goniometric measurements in ambulatory children with spastic cerebral palsy. Gait Posture 2000; 12: 114–121.
21. Eliasson A-C, Gordon A, Forssberg H. Basic co-ordination of manipulative forces of children with cerebral palsy. Dev Med Child Neurol 1991; 33: 661–670.
22. Lexell J, Downham D. How to assess the reliability of measurements in rehabilitation. Am J Phys Med Rehabil 2005; 84: 719–723.